

UNIVERSITY OF
ILLINOIS LIBRARY
AT URBANA-CHAMPAIGN
BOOKSTACKS

CENTRAL CIRCULATION BOOKSTACKS

The person charging this material is responsible for its renewal or its return to the library from which it was borrowed on or before the **Latest Date** stamped below. **The Minimum Fee for each Lost Book is \$50.00.**

Theft, mutilation, and underlining of books are reasons for disciplinary action and may result in dismissal from the University.

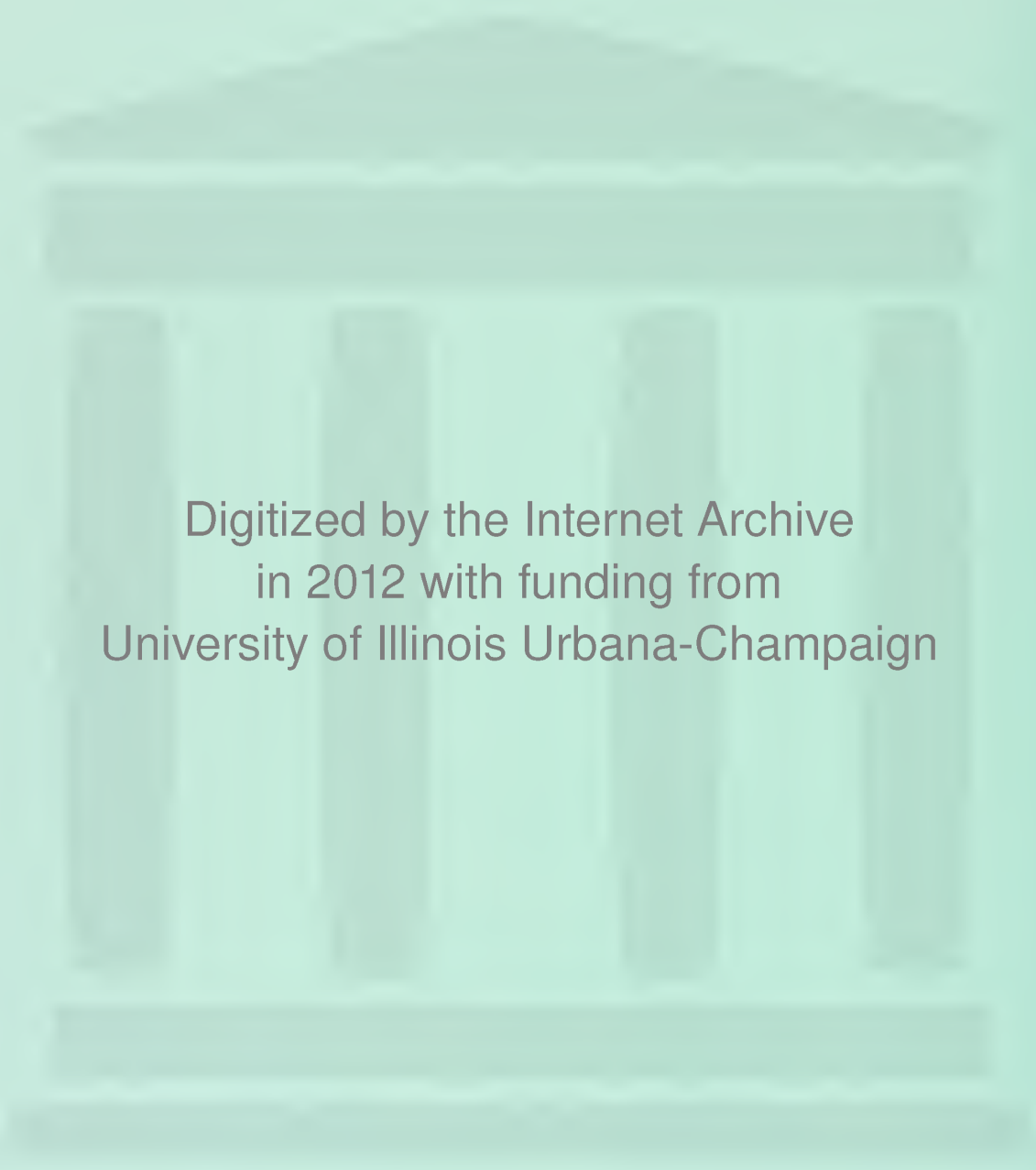
TO RENEW CALL TELEPHONE CENTER, 333-8400

UNIVERSITY OF ILLINOIS LIBRARY AT URBANA-CHAMPAIGN

MAY 04 1994

When renewing by phone, write new due date below
previous due date.

L162



Digitized by the Internet Archive
in 2012 with funding from
University of Illinois Urbana-Champaign

<http://www.archive.org/details/clarificationofr185joha>

Faculty Working Papers

A CLARIFICATION OF THE REDUNDANCY INDEX

Johny K. Johansson and Charles Lewis

#185

College of Commerce and Business Administration
University of Illinois at Urbana-Champaign

FACULTY WORKING PAPERS

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

May 31, 1974

A CLARIFICATION OF THE REDUNDANCY INDEX

Johny K. Johansson and Charles Lewis

#185

A CLARIFICATION OF THE REDUNDANCY INDEX

by

Johnny K. Johansson	and	Charles Lewis
Assistant Professor		Assistant Professor
Department of Business		Department of Psychology
Administration		

University of Illinois
Urbana, Illinois 61801

May 1974

ABSTRACT

The two main criticisms against Stewart and Love's redundancy index raised by Nicewander and Wood are examined. It is shown how Stewart and Love's original claim that the index represents the amount of overlapping or "redundant" variation between two sets of variables is justified. It is also shown why Nicewander and Wood's assertion that the redundancy index is not equal to the mean of the squared multiple correlations between a linear composite of one set of variables and the elements of the second set is incorrect.

Nicewander and Wood (1974) criticize the redundancy index $\overline{R_{y.x}^2}$ first proposed by Stewart and Love (1968). They dispute the two claims made by Stewart and Love that the $\overline{R_{y.x}^2}$ represents the proportion of variance of the variable set Y predictable from the variable set X, and that the index is the average of certain squared multiple correlations. This note intends to show that both claims are in fact justified.

Nicewander and Wood (hereafter NW) discuss the first claim by deriving the correlations between the original variables and their respective canonical variates. Thus, for the X set, these "loadings" are computed as

$$(1) \quad r_{xu_i} = R_{xx} c_i ,$$

and, for the Y set,

$$(2) \quad r_{yv_i} = R_{yy} d_i ,$$

where R_{xx} and R_{yy} represent the inter-correlation matrices of the X variables and the Y variables, respectively. The vectors c_i and d_i , $i=1,2,\dots,q_y$, are the eigenvectors whose elements are the weights determining the X-variates (or u_i) and the Y-variates (or v_i), respectively. There are q_y variables in the Y set.

To compute the redundancy index it is necessary to first compute the sum of the squared loadings. This sum becomes

$$(3) \quad r'_{xu_i} r_{xu_i} = c'_i R'_{xx} R_{xx} c_i = c'_i R_{xx}^2 c_i$$

for the X-set, and

$$(4) \quad r'_{yv_i} r_{yv_i} = d'_i R'_{yy} R_{yy} d_i = d'_i R_{yy}^2 d_i$$

for the Y-set. Stewart and Love (hereafter SL) define these sums as the variance extracted by the i 'th variate, $i=1,2,\dots,q_y$, from the X-set and Y-set, respectively. When the i 'th sum is divided by the number of variables in the set, "the resulting value is the proportion of the variance in the set extracted by that canonical variate" (SL,p.161).

NW argue that this terminology is misleading. First, they claim, all canonical variates have a variance of one since the usual constraints

$$(5) \quad c' R_{xx} c = d' R_{yy} d = 1$$

are in fact designed to insure this feature. To imply that variances differ is incorrect. Second, NW argue, the quantities derived in (3) and (4) are "empirically meaningless since both $c'_i R_{xx}^2 c_i$ and $d'_i R_{yy}^2 d_i$ are themselves devoid of any interpretation" (NW, p.93).

NW's first point seems to us to be of little relevance, and is possibly based upon a misreading of SL. Clearly, the equalities in

(3) and (4) are different from those of (5). The equations in (5) refer to the variances of certain variables; the equalities of (3) and (4) relate to the covariances between different variables.

The second criticism made by NW partly follows from the first and is thus also misplaced. In fact, one reason in favor of the SL approach is that a standard procedure in measuring the "explication" or "reproduction" or "extraction" by one variable or a linear combination of variables of the variance of another variable is to measure their correlation. The square of this correlation will then measure the percentage explanation obtained.

NW should be given some credit, however. When the summing over the squared variable loadings takes place, it should be kept in mind that these are not orthogonal for any one canonical variate, and thus speaking of a "total" amount of variation explained in the original variables becomes somewhat misleading. When the averaging over the number of variables is carried out, the resulting measure simply becomes the mean proportion of variance explained in each original variable by that canonical variate. However, with the original variables standardized, the variances all are one, so that even for the non-orthogonal case the SL statement quoted above (SL, p.161) is basically justified.

The second attack made by NW upon SL is somewhat more substantial. SL point out that their redundancy index $\overline{R_{y.x}^2}$ is equal to the mean squared multiple correlation, where the multiple correlations refer to

each Y-variable regressed upon the whole X-set. There is no proof of the assertion. NW first state that the SL presentation is not clear, and that there are two alternative interpretations of the assertion. They attempt to show that the assertion is incorrect under either interpretation.

Since NW's second interpretation represents a misunderstanding and is thus incorrect, we will here concentrate upon the first interpretation relating to each Y-variable regressed upon the X's. NW rewrites the squared multiple correlation between y_j , $j=1,2,\dots,q_y$, and an optimum linear combination of the X set as

$$(6) \quad R_{y_j \cdot X}^2 = r_{xy_j}' R_{xx}^{-1} r_{xy_j},$$

where r_{xy_j} is the j 'th column of R_{xy} , the matrix of intercorrelations between the X's and Y's. The authors then state: "Clearly, the average of these squared multiple correlations cannot be equal to the redundancy index $\overline{R_{y \cdot X}^2}$ " (NW, p.93). We will show here that this unproven assertion is in fact incorrect (although equation (6) is in itself correct).

Since the original SL assertion has not been rigorously proven before (although all empirical results indicate they are correct) it will be useful to fully develop the necessary algebraic relationships. In what follows we will first establish the correctness of the SL assertion for the case where the X and the Y matrices are of the same rank ($q_y = q_x = q$)

Then the generalization to different ranks, the number of X variables being greater than the number of Y variables ($q_x > q_y$), will be carried out.

Using NW's notation, the SL index of redundancy is calculated as

$$(7) \quad \overline{R_{y.x}^2} = \frac{1}{q} \sum_{i=1}^q \lambda_i^2 (r'_{yv_i} r_{yv_i}),$$

with the quantities defined as before, the λ_i denoting the i 'th canonical correlation. Since the exposition will be clearer using individual correlations, we note that

$$(8) \quad r'_{yv_i} r_{yv_i} = \sum_{j=1}^q r_{y_j v_i}^2.$$

To show that SL's assertion is true, we need to prove the following

Theorem: If $(X|Y)$ is a matrix of N observations on $2q$ variables with rank $2q$ (X being of order N by q , Y of order N by q), and $(U|V)$ are the corresponding canonical variates based on X and Y , respectively, then

$$(9) \quad \frac{1}{q} \sum_{i=1}^q \left(\sum_{j=1}^q r_{y_j v_i}^2 \right) \lambda_i^2 = \frac{1}{q} \sum_{j=1}^q R_{y_j \cdot X}^2.$$

Proof: We make use of the facts that

- 1) X and U are related by a non-singular transformation.
- 2) Similarly for Y and V .
- 3) $r_{u_i u_j} = r_{v_i v_j} = r_{u_i v_j} = 0$ for $i \neq j$, where $r_{j_i v_i} = \lambda_i$, $i, j = 1, 2, \dots$

For convenience, all variables are assumed standardized.

From 1) it follows that $R_{y_j \cdot X}^2 = R_{y_j \cdot U}^2$.

From 3) it follows that $R_{y_j \cdot U}^2 = \sum_{i=1}^q r_{y_j u_i}^2$

From 2) it follows that $y_j = \sum_{i=1}^q a_i v_i$, with

$$r_{y_j v_k} = \sum_{i=1}^q a_i r_{v_i v_k} = a_k, \text{ using 3).}$$

Similarly,

$$(10) \quad r_{y_j u_k} = \sum_{i=1}^q a_i r_{v_i u_k} \\ = a_k r_{v_k u_k}, \text{ again using 3).}$$

Combining these results, we have $R_{y_j \cdot X}^2 = R_{y_j \cdot U}^2 = \sum_{i=1}^q r_{y_j u_i}^2 = \sum_{i=1}^q a_i^2 r_{v_i u_i}^2$

$$(11) \quad R_{y_j \cdot X}^2 = \sum_{i=1}^q r_{y_j v_i}^2 r_{v_i u_i}^2 \\ = \sum_{i=1}^q r_{y_j v_i}^2 a_i^2.$$

Summing over j and dividing by q gives the desired result.

The generalization of this result to the case where the rank of the X matrix is q_X , $q_X > q_Y$, hinges on whether equality (10) is still valid. We will establish that it is by showing that the $r_{v_i u_k}$ vanish for $k = q_Y + 1, q_Y + 2, \dots, q_X$.

Again adopting NW's notation, we rewrite their equations (2) and (3) as

$$(12) \quad -\lambda R_{yy} d + R'_{xy} c = 0$$

$$(13) \quad R_{xy} d - \lambda R_{xx} c = 0.$$

This is the system of equations from which the canonical correlations are derived. As is well known, the system has a solution only if the determinant of the coefficients equals zero. This can be written as

$$(14) \quad \begin{vmatrix} -\lambda R_{yy} & R'_{xy} \\ R_{xy} & -\lambda R_{xx} \end{vmatrix} = 0.$$

The determinantal equation (14) forms a polynomial of degree $(q_y + q_x)$ in λ . The positive roots of this polynomial, in descending order, yield the canonical correlations. We will show that there are at least $(q_x - q_y)$ zero roots, and that there are q_y nonnegative and q_y nonpositive roots. Of prime interest in canonical analysis are the q_y nonnegative roots, generating, as NW indicate, the canonical correlations λ_i , with corresponding vectors c_i and d_i , $i=1,2,\dots,q_y$.

Relying on a well known result on the determinant of a partitioned matrix (see, e.g., Dhrymes, 1970, p. 570), we can write (14) as

$$(15) \quad \begin{vmatrix} -\lambda R_{xx} & \\ & -\lambda R_{yy} - R'_{xy} (-\lambda R_{xx})^{-1} R_{xy} \end{vmatrix} = 0.$$

The validity of the result requires $|\lambda R_{xx}| \neq 0$ which holds unless there is an exact linear relationship between some X-variables. By factoring out the appropriate terms, we see that (15) can be written

$$(16) \quad (-\lambda)^{q_x} |R_{xx}| \left| -\frac{1}{\lambda} (\lambda^2 R_{yy} - R_{xy}' R_{xx}^{-1} R_{xy}) \right| = 0,$$

and, further,

$$(17) \quad (-1)^{q_x+q_y} \lambda^{q_x-q_y} |R_{xx}| \left| \lambda^2 R_{yy} - R_{xy}' R_{xx}^{-1} R_{xy} \right| = 0.$$

Here use is made of the property that, for any constant μ and any non-singular matrix A of rank m , $|\mu A| = \mu^m |A|$. From (17) we see that the determinantal equation is satisfied for $(q_x - q_y)$ zero roots of λ . The nonzero roots yielding the usual canonical correlations are in fact the nonzero roots of

$$(18) \quad \left| \lambda^2 R_{yy} - R_{xy}' R_{xx}^{-1} R_{xy} \right| = 0.$$

Thus, the complete set of correlations are made up of the q_y roots extracted from (18) augmented by the $(q_x - q_y)$ zero roots from (17) and we have the desired result:

$$(19) \quad r_{v_i u_k} = 0, \text{ for all } k = q_y + 1, q_y + 2, \dots, q_x.$$

Accordingly, we can write in the general case

$$\begin{aligned}
 (20) \quad R_{y_j \cdot U}^2 &= \sum_{i=1}^{q_x} r_{y_j u_i}^2 \\
 &= \sum_{i=1}^{q_y} r_{y_j u_i}^2
 \end{aligned}$$

since, from (19),

$$(21) \quad r_{y_j u_k} = \sum_{i=1}^{q_y} a_i r_{v_i u_k} = 0, \text{ for } q_x \geq k > q_y.$$

The proof for the same rank case ($q_y = q_x = q$) can then be applied directly to the general case ($q_y < q_x$).

It deserves to be emphasized, that this result is not immediately obvious. In words, it means that the canonical variates comprise the total amount of variation in the X-variables which is relevant to the original variation in the Y-set. All residual variation in the X-set is orthogonal to the original Y-variables.

REFERENCES

Dhrymes, Phoebus J., Econometrics: Statistical Foundations and Application. New York, Harper and Row, 1970.

Nicewander, W. Alan, and Donald A. Wood, "Comments on 'A General Canonical Correlation Index' ", Psychological Bulletin, 1974, Vol.81, No.1, 92-94.

Stewart, Douglas, and William Love, "A General Canonical Correlation Index", Psychological Bulletin, 1968, Vol.70, No.3, 160-163.

Van de Geer, John P., Introduction to Multivariate Analysis for the Social Sciences. San Francisco: W. H. Freeman and Company, 1971.



